Corpus-Based Associations Provide Additional Morphological Variants to Medical Terminologies

Pierre Zweigenbaum, Ph.D. Natalia Grabar, M.Sc.

Mission de recherche en Sciences et Technologies de l'Information Médicale, STIM/DPA/DSI, Assistance Publique – Hôpitaux de Paris & ERM 202, INSERM, France

{pz,ngr}@biomath.jussieu.fr http://www.biomath.jussieu.fr/~{pz,ngr}/

Knowledge of morphologically derived words, as provided for medical English by the UMLS Specialist Lexicon, is useful to detect term variants for automated coding and indexing. For most other languages though, no comparable morphological knowledge base is available. We therefore endeavored to design general methods to help collect such knowledge for a given language. We propose here a method for discovering derived words in text corpora and apply it to a French medical corpus. To evaluate this method, we study its ability to suggest derived adjectives for 2,297 nouns found in the SNOMED nomenclature, which itself specifies adjectival equivalents for some of its terms. 74% of the proposed adjectives are judged correct (precision) and cover 16% of these nouns (recall), a larger amount than what SNOMED already specifies. Furthermore, the corpus suggests additional adjectives which can increase SNOMED's by 76%. We conclude that such a method can help speed up the construction of a morphological knowledge base which can increase the number of term variants in an existing controlled vocabulary.

INTRODUCTION

Controlled vocabularies associate with each concept of a given domain a preferred term,[1] often complemented with additional, equivalent terms, called 'synonyms' or (*e.g.*, in MeSH®) 'entry terms'. These variant expressions of the same concept often use morphologically related words. For instance, the UMLS Metathesaurus® 2002AA compiled 16 different terms for concept C0027051, *Myocardial Infarction*, in which *infarction* / *infarct* / *infarcts* (and *myocardial* / *myocardium*) form a morphological family. Inflection (*infarct* / *infarcts*), derivation (*infarct* / *infarction*; *myocardium* / *myocardial*) and compounding (*myocardium* / *myocardiopathy* / *encephalomyocarditis*) are classical distinctions[1] among these morphological relations. Knowledge of such relations is instrumental in the automatic detection of term variants,[2,3] which is itself a key component for automated indexing.[4,5] Such knowledge is provided for medical English by the UMLS Specialist Lexicon®.[6] Several teams have worked on morphological knowledge for general[7] or medical[8,9] French, and are now preparing a unified medical lexicon for French (UMLF[10]).

We have shown in earlier work how structured terminologies can be exploited to learn morphological variants for various languages;[9,11] such discovery techniques will help to compile the morphological relations in the UMLF lexicon. Controlled vocabularies are an interesting source for morphological discovery since they concentrate a high density of specialized vocabulary which they link through numerous semantic relations (synonymy, hypernymy, etc.). However, they cannot mirror all actual term usage: large, diversified corpora of medical texts[12] can provide complementary help in this purpose. The general goal of the present work is to provide automated methods to assist the collection of morphological knowledge from corpora.

Various kinds of corpus-based, morphological discovery methods have been proposed. Jacquemin[3] matches two-word MeSH terms with variant digrams (two-word sequences) in a corpus. Xu and Croft[13] filter the morphological variants found by Porter's stemmer[14] in a corpus according to their association strength in that corpus. We have designed a method[15,16] which blends together elements from Xu and Croft's method and from our earlier terminology-based work.[9] It only needs a corpus, and is not restricted to two-word terms. The specific goal of the present work is to evaluate its ability to add new morphological variants to terms in a controlled vocabulary. We chose SNOMED®, more precisely the French SNOMED Microglossary for Pathology,[17] as a test bed, because it combines several distinct features: it uses mixed-case, accented letters, and is therefore directly usable for natural language processing; as a multiaxial terminology, it provides terms in several semantic axes (Topography, Morphology, etc.), allowing us to work on diversified words; and among its rich set of synonyms, it explicitly specifies adjectival equivalents (class '05') to nominal terms (*e.g.*, *foetus* / *foetal*). We therefore focus on nouns and their derived adjectives, and investigate (i) whether the morphologically related words (noun / adjective) explicitly specified by SNOMED can be found by our corpus-based method, and (ii) whether new morphologically related words can add to SNOMED's existing variants. The method is tested on SNOMED as a kind of formal proof of concept, but it will be all the more useful on thesauri which lack such informa-

tion, such as MeSH, ICD, ICF, etc.

We first describe the corpus which we shall exploit and the test set of nouns and adjectives obtained from SNOMED. We then summarize how pairs of nouns and derived adjectives are detected in the corpus and how we evaluate their contribution with respect to SNOMED, and present and discuss evaluation results.

MATERIAL

The corpus used in this experiment was initially compiled for working on cross-language information retrieval.[18] It takes advantage of the CISMeF catalog (www.chu-rouen.fr/cismef/[19]) which indexes French-language medical Web sites with MeSH keywords. We downloaded all HTML pages indexed under the main heading *Signs and Symptoms* (C23) or one of its descendants, extended to pages found one hyperlink farther below these initial pages. 4,627 pages were obtained and converted to plain text; the language of each line was identified, and lines written in languages other than French were filtered out. Finally, each word was tagged with its part-of-speech (with TreeTagger[20]) and lemmatized (with FLEMM[21]). Grammatical words were discarded, leaving 2,041,627 (lemmatized) tokens of categories noun, adjective, verb or adverb.

The French Microglossary for Pathology[17] is a subset of the full SNOMED International; it contains 12,550 terms, which we tagged and lemmatized in earlier work.[11] To build a test set of medical nouns, we collected all of its terms which consist of a single noun, possibly followed by the expression *', SAI'* (French acronym for *not otherwise specified*). This provided a test set of 2,297 nouns. As mentioned in the introduction, SNOMED explicitly specifies adjectival equivalents for some of its concepts. We thus collected in a similar way all terms consisting of a single adjective. We also obtained the associations of such nouns and adjectives: those linked with the same SNOMED code. When several adjectives were linked to the same noun(s), we manually selected the appropriate morphological association(s) (*e.g.*, concept *M-14400* has nouns *déchirure*, *lacération* and *rupture*, and adjectives *déchiré*, *lacéré* and *rupturé*, which we associated in three noun-adjective pairs). 435 associations were obtained for 345 different nouns and constitute our SNOMED reference.

METHODS

Corpus-Based Discovery of Derived Words

Our method for discovering morphologically related words in a corpus, just like our previous work with structured terminologies,[9] looks for words that (i) have a similar form and (ii) occur in a semantically constrained context. Our simple test for formal similarity checks whether two words start with the same C characters (C = 4 in the present experiment). We also observe, following Xu and Croft,[13] that for reasons of thematic continuity in a text, semantically related words are often found together at a moderate distance from each other. To detect these cooccurrences, we scan the words in the corpus through an N-word sliding window and count the cooccurrences of each pair of formally similar words. The log-likelihood ratio[22] is used to measure in a more principled way the degree of (in)dependence, hence the association strength of two such words. In summary, our principle is that words which share the same first C characters and which cooccur in an N-word window more often than chance are likely to be morphologically related (N is set to 150 in this work). This indeed collects both words which are morphologically related (through derivation and compounding) and some amount of spurious associations. In the present experiment, we want to focus on adjectives derived from nouns. We therefore added three filters. A pair of words collected by the method is considered to be produced by a hypothetical rule which substitutes their suffixes (*e.g., abdom<u>en</u> / abdom<u>inal</u>*). A formal filter tests the sizes of these suffixes: the derived word must be longer than its base word (to allow for some flexibility, we accept one character less than the base), but not more than five characters longer so as to block some compounds; and the total length of both suffixes must not exceed ten characters (*e.g., adénom<u>atose</u> / adénoc<u>arcinome</u>* is blocked). Of course, we select the category of the base and the derived word (N and A). Finally, we take into account the frequency of application of the 'rule' on our test data: a rule which applies only once is likely to denote a spurious cooccurrence (*e.g., calib<u>re</u> / calic<u>iel</u>*), unless this cooccurrence has a very high association strength. We experimentally set an association threshold of 50 under which rules which apply only once are discarded (with the word pairs they would produce).

Evaluation

We first evaluate the intrinsic propensity of the method to propose and order pairs of nouns and derived adjectives found in the corpus. We reviewed each pair produced from the CISMeF corpus and computed the ratio of correct pairs (actual derived adjectives) over the total number of pairs considered. Considering word pairs in decreasing order of association strength, we compute a *cumulated precision* on all pairs seen up to a given rank and a *local precision* on successive slices of 200 word pairs. We then evaluate the adjectives proposed for the 2,297 nouns of our test set. We compute their specific *precision* (#correct / #proposed) and the ratio of correct associations over the

total number of nouns in the test set (*recall*). We also compare the proposed adjectives with those specified by SNOMED, compute a recall relative to SNOMED, and also the proportion of correct, new derived adjectives added by the corpus-based method with respect to SNOMED. Programs were implemented using Perl and Unix scripts and PostgreSQL queries.

RESULTS

5,036 derived noun-adjective pairs are selected by the system in the corpus. Cumulated and local precision are plotted on figure 1. The overall precision is moderate (77% cumulated precision at highest rank); a high association strength generally favors a larger proportion of correct derived pairs (figure 1a), although some variation is observed in local precision. A low rule frequency is generally a clue of an incorrect pair (figure 1b); the impact of a high rule frequency is less clear. Note that this set of derived noun-adjective pairs is obtained from a larger set of 14,463 candidate derived pairs, themselves selected from 48,003 corpus associations, 44% of which are correct.[16]

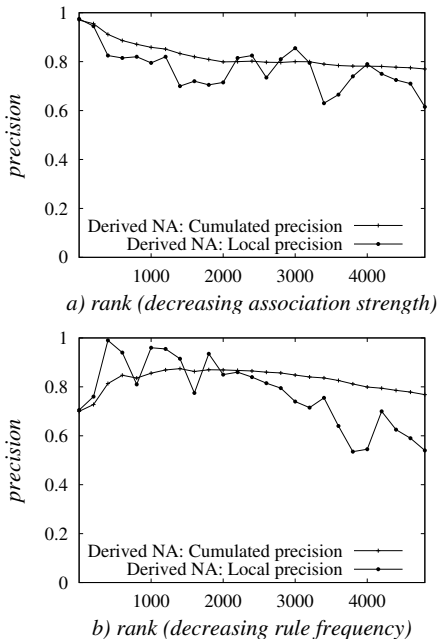

Figure 1: Cumulated and local precision of selected cooccurring candidate noun-adjective derived pairs, plotted against rank: a) by decreasing order of association strength; b) by decreasing order of rule frequency.

Table 1 shows an excerpt of candidate noun-adjective pairs. Asterisks mark associations of unrelated (or too far away) terms; question marks tag related words whose relevance is questionable in our context.

Table 2 shows the amount of nouns and associated adjectives obtained from SNOMED and selected by our corpus-based method. Their distribution over SNOMED axes is skewed: SNOMED provides noun-adjective associations mainly for the M and T axes, and many G preferred terms are themselves adjectives. The corpus method proposes 502 noun-adjective associations, 370 of which are correct (precision = 74%, recall = 16%). They cover 31% of SNOMED-specified noun-adjective associations, and propose 263 (relative +76%) new derived adjectives. Not shown on this table, while 48% of these adjectives occur elsewhere within SNOMED terms, 52% do not occur at all in the French Microglossary for Pathology; instances are *apoptotique*, *bacillaire*, *bronchiteux*.

Table 1: Candidate derived noun-adjective pairs: first group with highest association strength, second group with lowest. When the noun is in the SNOMED-derived test set, the initial of the SNOMED axis is shown in parentheses. When the adjective is specified by SNOMED for this noun, it is followed by a + sign.

diabète / *diabétique*; *(D) asthme* / *asthmatique*; *(T) urine* / *urinaire*; *(T) cellule* / *cellulaire+*; *(M) kyste* / *kystique*; *douleur* / *douloureux*; *(D) tuberculose* / *tuberculeux*; *grippe* / *grippal*; *(M) cancer* / *cancéreux*; *vaccin* / *vaccinal*; *glomérule* / *glomérulaire*
format / *formateur**; *réserve* / *réservé?*; *réalité* / *réalisable?*; *utilité* / *utilisable?*; *forme* / *formateur**; *signal* / *signalé?*; *clinical** / *clinique*; *(P) consultation* / *consultable?*

DISCUSSION

The present method for corpus-based discovery of noun-adjective associations can help to find derived adjectives for 16% of the nouns in our SNOMED-derived test set (*i.e.*, 16% global recall). Although this is a rather low figure, it is comparable with what SNOMED explicitly specifies (15% recall). Furthermore, this proposes new derived adjectives to complement SNOMED's, affording it derived adjectives for 26% of the nouns in our test set instead of the current 15%, a potential increase of 11% in SNOMED's recall (relative increase of 76%). Indeed, these derived adjectives need to be further reviewed by the SNOMED editors who alone can judge their final relevance with respect to SNOMED's organizational principles. Human editors may have very good reasons to make different local choices. For instance, in the French SNOMED, adjective *urinaire* (*urinary*) is associated with *voies urinaires* (*urinary tract*), and is therefore not considered an adjectival equivalent of *urine*. This raises the issue of what should be an 'acceptable' derived adjective in a medically-oriented lexicon such as UMLF. We believe a reasonable position is to com-

Table 2: Nouns and associated adjectives in French SNOMED Microglossary for Pathology, and candidate derived adjectives selected from corpus cooccurrence data.

	Axis	A	C	D	F	G	J	L	M	P	S	T	Total
Snomed	Nouns	48	156	367	187	25	3	313	773	43	5	377	2297
	Adjectives				7	378			165			170	720
	Associations (#nouns)				6	5			173			161	345
Corpus	Corpus associations	10	34	98	45	8	2	16	133	3		153	502
	Correct	6	25	49	33	5	1	10	97	3		141	370
	Correct and new	6	25	49	33	4	1	10	69	3		63	263
	Global precision	60%	74%	50%	73%	62%	50%	62%	73%	100%		92%	74%
	Global recall	12%	16%	13%	18%	20%	33%	3%	13%	7%		37%	16%
	Addition % global	12%	16%	13%	18%	16%	33%	3%	9%	7%		17%	11%
	Recall % SNOMED				0	20%			16%			48%	31%
	Addition % SNOMED				550%	80%			40%			39%	76%

pile the various linguistically relevant derived adjectives in the lexicon and to let lexicon users select those that suit their needs. A finer semantic categorization of different derivation operators, as is usually made in more linguistically-oriented work,[7] may also help make such decisions. For instance, relational adjectives (*pertainyms*,[1] *e.g.*, *diabétique*, *vaccinal*) have a simple, general semantics of *related to X* where *X* is the base noun. In contrast, *-able* adjectives (*e.g.*, *consultable*) imply a more specific semantic relation (*e.g.*, *X-able* means *which may be X-ed*, where *X* is a base verb), which may motivate a different decision. We examined the reasons for not finding more derived adjectives. Some are intrinsic limitations of the method (four-character threshold, finding both noun and adjective in the selected corpus, etc.). We also found that many of the nouns in our test set simply do not have attested derived adjectives (*e.g.*, *avorton* – English *abortus*, *dermatite*, *hidrosadénite*, *éphélide*, *aortite*), even through manual search with the Google® or AltaVista® search engines. This means that our measurement underestimates the actual recall of the method. This silence can be reduced in several ways. Indeed, enlarging and diversifying the corpus will accumulate new words: less than 5,000 documents were used, many times more can be easily collected, both on the Web and in patient files. Using the induced rules to collect more word pairs, as we did on terminologies,[9] is another way of extending coverage. Finally, lowering thresholds, in particular the 4-character common initial substring length, increases recall but also decreases precision. There will also always remain a need for linguistic knowledge which requires human intervention. This is the case, for instance, of suppletive bases (*e.g.*, the base *card-* for words derived from *heart*). This being said, the present, standalone evaluation shows that the proposed method can detect new, morphologically-related word pairs. We believe it does has value, not in isolation, but as one of a series of complementary methods (terminology-based,[9] terminology+corpus,[3] etc.) for collecting good-quality morphological knowledge, which can be combined to obtain both better recall and precision.

The corpus-selected adjectives obtain an average precision of 74%, which we consider acceptable since human reviewers can quickly eliminate the 26% remaining errors. Improving this precision is nevertheless a theme of current work. Errors include actual derived adjectives which were not the expected relational adjective: *figure* / *figuré* (*facial*), *embryon* / *embryonné* (*embryonnaire*), *travail* / *travailleur*; *facial* is built with a 'suppletive' base (*face* or *facies* instead of *figure*), which cannot be detected by our method. Words related through several derivation steps were also considered as errors: *dentier* / *dentaire*, *colite* / *colique*, *conjonctivite* / *conjonctival*. Some neoclassical compounds passed our heuristic selection criteria: *hydrocèle* / *hydrique*, *insecte* / *insecticide*. Some words were incorrectly tagged as adjectives: nonwords (*côlon* / *côlonb*, *muscle* / *musclaire*), foreign words which passed our language filter (English *contraction* / *contracted*, Spanish *complication* / *complicatione(s))* or actual French words (*cornée* / *corné*). Simple coincidences were also encountered, such as *collapsus* / *collatéral*. Some of the discovered 'rules' cause an important proportion of the errors, *e.g.*, the rule which adds a final *-e* is a common French inflection rule, but not a usual derivation rule. Progressive inclusion of knowledge to validate or invalidate some of the discovered rules[7] will help increase further the current precision when tackling new corpora.

Let us note finally that the method, given initial part-of-speech tagged and lemmatized data, does not need language-specific knowledge beyond the suffix-length thresholds. These may need small adjustments for languages with much longer (or shorter) suffixes. Among the proposed adjectives, some are specific to medical usage (*e.g.*, *paramètre* / *paramétrial* instead of *paramétrique* – English *parametrium* / *parametrial* or *parametric*) and are more likely to be absent from gen-

eral language resources: an advantage of corpora is that they reveal actual word usage, which is more difficult for controlled vocabularies to keep up with.

CONCLUSION

The present experiment shows the potential of text corpora as a source of morphological knowledge. The main contribution of the present kind of work is to design means for automatically analyzing large amounts of data (corpora) and proposing a synthesis of candidate knowledge elements extracted from that data. Its rationale, which is implemented in the UMLF project, is to prepare data for human editors so as to help them do faster, more complete work. We also believe that the present method, tested here on French SNOMED, should be effective on other languages and thesauri.

Acknowledgments

We thank again Dr. R.A. Côté (French SNOMED Microglossary for Pathology), S.J. Darmoni and the CISMeF team (CISMeF), H. Schmid (TreeTagger) and F. Namer (FLEMM). F. Hadouche implemented the corpus cooccurrence processing program.

REFERENCES

1. McCray AT. The nature of lexical knowledge. *Methods Inf Med* 1998;37(4–5):353–60.
2. Aronson AR. Effective mapping of biomedical text to the UMLS Metathesaurus: The MetaMap program. *J Am Med Inform Assoc* 2001;8(suppl).
3. Jacquemin C. Guessing morphology from terms and corpora. In: Proc 20th ACM SIGIR, Philadelphia, PA. 1997:156–67.
4. Aronson AR, Bodenreider O, Chang F, et al. The NLM indexing initiative. *J Am Med Inform Assoc* 2000;7(suppl):17–21.
5. Hahn U, Honeck M, Piotrowski M, and Schulz S. Subword segmentation: Leveling out morphological variations for medical document retrieval. *J Am Med Inform Assoc* 2001;8(suppl):229–33.
6. McCray AT, Srinivasan S, and Browne AC. Lexical methods for managing variation in biomedical terminologies. In: Proc 18th Annu Symp Comput Appl Med Care, Washington. Mc Graw Hill, 1994:235–9.
7. Hathout N, Namer F, and Dal G. An experimental constructional database: the MorTAL project. In: Boucher P, ed, *Many morphologies*. Cascadilla Press, Somerville, MA, 2002:178–209.
8. Lovis C, Baud R, Rassinoux AM, Michel PA, and Scherrer JR. Medical dictionaries for patient encoding systems: a methodology. *Artif Intell Med* 1998;14:201–14.
9. Grabar N and Zweigenbaum P. Language-independent automatic acquisition of morphological knowledge from synonym pairs. *J Am Med Inform Assoc* 1999;6(suppl):77–81.
10. Zweigenbaum P, Baud R, Burgun A, et al. Towards a unified medical lexicon for French. In: Baud R, Fieschi M, Le Beux P, and Ruch P, eds, Proc Medical Informatics Europe, Amsterdam. IOS Press, 2003:415–20.
11. Grabar N and Zweigenbaum P. Automatic acquisition of domain-specific morphological resources from thesauri. In: Proc RIAO 2000, Paris, France. C.I.D., April 2000:765–84.
12. Zweigenbaum P, Jacquemart P, Grabar N, and Habert B. Building a text corpus for representing the variety of medical language. In: Patel VL, Rogers R, and Haux R, eds, Medinfo, 2001.
13. Xu J and Croft BW. Corpus-based stemming using co-occurrence of word variants. *ACM Transactions on Information Systems* 1998;16(1):61–81.
14. Porter MF. An algorithm for suffix stripping. *Program* 1980;14:130–7.
15. Hadouche F. Acquisition de resources morphologiques à partir de corpus. DESS d'ingénierie multilingue, Institut National des Langues et Civilisations Orientales, Paris, 2002.
16. Zweigenbaum P, Hadouche F, and Grabar N. Apprentissage de relations morphologiques en corpus. In: Daille B, ed, Proc TALN 2003 (Traitement automatique des langues naturelles), Batz-sur-mer. ATALA, IRIN, June 2003:285–94.
17. Côté RA. Répertoire d'anatomopathologie de la SNOMED internationale, v3.4. Université de Sherbrooke, Sherbrooke, Québec, 1996.
18. Chiao YC and Zweigenbaum P. Looking for French-English translations in comparable medical corpora. *J Am Med Inform Assoc* 2002;8(suppl):150–4.
19. Darmoni SJ, Leroy JP, Thirion B, et al. CISMeF: a structured health resource guide. *Methods Inf Med* 2000;39(1):30–5.
20. Schmid H. Probabilistic part-of-speech tagging using decision trees. In: Proceedings of the International Conference on New Methods in Language Processing, Manchester, UK. 1994:44–9.
21. Namer F. FLEMM : un analyseur flexionnel du français à base de règles. *Traitement Automatique des Langues* 2000;41(2):523–47.
22. Manning CD and Schütze H. *Foundations of Statistical Natural Language Processing*. MIT Press, Cambridge, MA, 1999.